Calibrated Regression-as-Classification for Probabilistic Forecasting

Jef Jonkers
IDLab
Department of Electronics and Information Systems
Ghent University
Belgium

Glenn Van Wallendael
IDLab
Department of Electronics and Information Systems
Ghent University - imec
Belgium

Luc Duchateau
Biometrics Research Group
Department of Morphology, Imaging, Orthopedics, Rehabilitation and Nutrition
Ghent University
Belgium

Sofie Van Hoecke
IDLab
Department of Electronics and Information Systems
Ghent University - imec
Belgium

Abstract

Regression-as-classification (R2C) is a pragmatic approach to distributional regression in which the response space is discretized, bin membership probabilities are estimated, and a monotone discrete cumulative distribution function (CDF) is subsequently reconstructed. The remaining challenge is *calibration*, ensuring predicted probabilities align with empirical frequencies in finite samples. We propose post-hoc calibration procedures for R2C predictors that are *conformal predictive systems* (CPS). These systems produce predictive CDF bands that, under the assumption of exchangeability, are guaranteed to enclose the out-of-sample calibrated distribution. Specifically, we propose four calibration methods that respectively target probabilistic calibration, grid calibration (a notion we introduce), isotonic calibration, and autocalibration, all of which are compatible with discrete CDF outputs.

1 Introduction

Modern machine learning applications increasingly demand probabilistic forecasts instead of single-point estimates: areas such as clinical risk assessment, weather prediction, finance, and policy rely on predictive distributions that can capture heteroskedasticity, heavy tails, and multi-modal behavior, while remaining *calibrated*, i.e., ensuring that the model gives plausible probabilistic explanations of observations (Gneiting and Katzfuss, 2014; Gneiting and Resin, 2023). In distributional regression, the aim is to learn the conditional distribution $Y \mid X = x$, commonly represented via a predictive cumulative distribution function (CDF) $\hat{F}(\cdot \mid x)$ (or equivalent formulations such as quantile functions or probability densities).

A widely used and practical approach is to recast *regression as classification* by discretizing the response domain and predicting probabilities over bins or thresholds. This "regression-as-classification" (R2C) paradigm appears under many names, i.e., binning/quantization, ordinal or threshold regression, and digitized density estimation (Foresi et al., 1995; Weigend and Srivastava, 1995; Wilks, 2009; Messner et al., 2014; Fu et al., 2018; Tansey et al., 2016; Li et al., 2021; Chen et al., 2022; Stewart et al., 2023; Pintea et al., 2023; Zhang et al., 2016; van den Oord et al., 2016). R2C is popular because (i) classification losses are stable and well-understood (Fu et al., 2018; Stewart et al., 2023; Pintea et al., 2023), (ii) the output is naturally bounded in $[0, 1]$ (Foresi et al., 1995; Wilks, 2009), (iii) complex conditional shapes (e.g., multimodality) can be represented with sufficient resolution (Weigend and Srivastava, 1995; Tansey et al., 2016; Guha et al., 2023), and (iv) the resulting discrete distribution can provide a convenient interface for post-hoc calibration, such a conformalization (Guha et al., 2023; Jonkers et al., 2026). Early instances include fitting binary models at multiple thresholds to approximate a conditional CDF (Foresi et al., 1995) and fractional/soft binning schemes (Weigend and Srivastava, 1995). More recent work develops deep density estimators via multinomial heads (Tansey et al., 2016; Chen et al., 2022), randomized partitions and joint binary cross-entropy objectives (Berg et al., 2021; Li et al., 2021), and conformal post-processing tailored to R2C predictors (Guha et al., 2023; Jonkers et al., 2026).

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

In this work, we study R2C forecasters that output a valid discrete predictive CDF on a fixed grid, and we address the key remaining challenge of finite-sample calibration. Building on conformal predictive systems (CPS) (Vovk et al., 2019; Allen et al., 2025), we derive post-hoc calibration procedures that produce predictive CDF bands with finite-sample calibration guarantees under exchangeability. We instantiate this framework with four practical calibrators that target different notions of calibration (probabilistic calibration, grid calibration (introduced here), isotonic calibration, and auto-calibration), while remaining compatible with discrete CDF outputs.

2 Regression-as-classification for discrete CDF forecasting

Our main contribution is four post-hoc calibrators as CPS with finite-sample calibration guarantees (Section 3) for R2C distributional regressors. This section defines a simple, stable, and expressive base forecaster: a valid discrete CDF obtained from a softmax probability mass function (PMF) head and trained with joint binary cross-entropy (JBCE), which is a strictly proper scoring rule for conditional threshold probabilities on a chosen grid.

2.1 Valid CDF via a PMF head

Fix thresholds $T = \{t_0 < t_1 < \ldots < t_m < t_{m+1}\}$, with $t_0 := -\infty$ and $t_{m+1} := +\infty$ as boundary values, defining bins $I_l := (t_{l-1}, t_l]$ for $l = 1, \ldots, m+1$. Let $p_\theta(x) \in \Delta^{m+1}$ be the softmax output, interpreted as

$$p_{\theta,l}(x) = \mathbb{P}_\theta\big(Y \in I_l \mid X = x\big), \qquad l = 1, \ldots, m+1.$$

This PMF induces grid CDF values on T by cumulative summation, for $k = 1, \ldots, m$:

$$F_{\theta,k}(x) := \sum_{j=1}^{k} p_{\theta,j}(x) = \mathbb{P}_\theta\big(Y \le t_k \mid X = x\big),$$

so $0 \le F_{\theta,1}(x) \le \cdots \le F_{\theta,m}(x) \le F_{\theta,m+1}(x) = 1$ by construction.

2.2 Training with JBCE

For each threshold $t_k \in T$, define labels $z_{ik} := \mathbb{1}\{Y_i \le t_k\}$. With Bernoulli log-loss $\ell(z, q) := -z \log q - (1 - z) \log(1 - q)$, we train by minimizing JBCE (Li et al., 2021):

$$\widehat{\mathcal{R}}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} \ell\big(z_{ik},\, F_{\theta,k}(X_i)\big). \tag{1}$$

Let $F_k^*(x) := \mathbb{P}(Y \le t_k \mid X = x)$ denote the true conditional CDF values on the grid. JBCE is strictly proper (Gneiting and Raftery, 2007) for the vector $(F_1^*(x), \ldots, F_m^*(x))$ since Bernoulli log-loss is uniquely minimized at the true mean, and summation over k preserves strictness (our monotone parameterization imposes a structured model class).

3 Finite-sample calibration guarantees

We now provide the discrete CDF forecaster from Section 2 with finite-sample, out-of-sample calibration guarantees under exchangeability by interpreting calibration through the framework of conformal predictive systems (CPS) (Vovk et al., 2019; Allen et al., 2025). Because achieving exact out-of-sample calibration for a single predictive distribution is generally impossible without imposing strong assumptions, CPS instead outputs a credal set of probabilistic forecasts that, under weak symmetry conditions, is guaranteed to include at least one calibrated predictive distribution. For real-valued responses, this credal set can be encoded as a predictive system, that is, a band of CDFs $\Pi = \{(y, u) : \Pi_\ell(y) \le u \le \Pi_u(y)\}$, where the lower and upper envelopes Π_ℓ and Π_u are both monotone increasing. The thickness of the bounds can be seen as a quantification of epistemic uncertainty (Allen et al., 2025).

3.1 Predictive system construction from in-sample calibration

Let $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ be exchangeable, and write the empirical measure $\mu := \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ on $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$. Following Allen et al. (2025), consider any distributional regression procedure

$$G :\ (\mu, x, y) \mapsto G[\mu, x](y) \in [0, 1],$$

such that $G[\mu, x](\cdot)$ is a CDF.

Define the associated predictive system bounds at test covariate X_{n+1} by

$$\Pi_\ell[\mu, X_{n+1}](y) := \inf_{y' \in \mathbb{R}} G[\mu + \delta_{(X_{n+1}, y')},\, X_{n+1}](y), \tag{2}$$

$$\Pi_u[\mu, X_{n+1}](y) := \sup_{y' \in \mathbb{R}} G[\mu + \delta_{(X_{n+1}, y')},\, X_{n+1}](y). \tag{3}$$

Intuitively, we augment the sample with the test covariate X_{n+1} paired with every possible label y', run the in-sample calibrated procedure G, and retain only the pointwise envelope in y.

The central result of Allen et al. (2025) states that if G is *in-sample calibrated* (conditionally on μ), then the predictive system $\Pi[\mu, X_{n+1}]$ contains an *out-of-sample calibrated* predictive CDF for Y_{n+1} (with the

calibration notion inherited from G). In particular: (i) probabilistically calibrated G yields a predictive system that contains a probabilistically calibrated forecast (probability integral transform (PIT, $F(Y)$)-uniformity in the generalized sense); (ii) isotonically calibrated G yields an isotonically calibrated forecast (hence also threshold/quantile/PIT calibration under mild conditions); (iii) auto-calibrated G yields a predictive system containing an auto-calibrated forecast.

In practice, we use a split setup with an estimation set D_0 to train the base model (and any summaries), and a calibration set D_1 (size n_1) to form μ and compute the associated predictive system bounds (2)-(3). The conformal calibration guarantees remain finite-sample valid under exchangeability of $(D_1, (X_{n+1}, Y_{n+1}))$ conditional on D_0 (Allen et al., 2025).

3.2 Conformal calibrators for discrete CDFs

We now specify G (and a covariate summary H when needed) for four practical calibration notions (probabilistic, grid, isotonic, and auto-calibration). All guarantees below are finite-sample under exchangeability.

Probabilistic calibration: CM-CPS. We use a CPS defined by a conformity measure (CM-CPS) based on the predicted CDF. Concretely, following Vovk et al. (2020) and Jonkers et al. (2025), we use the PIT $S(x, y) = \hat{F}(y \mid x)$, which is one of the few conformity measures that preserve universal approximation when the underlying base learner itself is a universal approximator. The resulting predictive system Π^{CM} yields a *probabilistically calibrated* predictive CDF for Y_{n+1}, which in turn guarantees valid marginal coverage for all associated prediction intervals (Vovk et al., 2019; Allen et al., 2025).

Auto-calibration: CBIN. We use conformal binning (CBIN) (Allen et al., 2025), a CPS obtained by applying conformalization to a binning-based distributional regressor. We define a summary $H(x)$ of the base predictive distribution (here the estimated CDF) and build bins $B(\cdot)$ on $H(X)$ using D_0 (e.g. k-means). On D_1, the distributional regression procedure G^{BIN} returns the empirical CDF of labels in the bin of x:

$$G^{\text{BIN}}[\mu, x](y) = \frac{1}{|I(x)|} \sum_{i \in I(x)} \mathbb{1}\{Y_i \le y\}.$$

where $I(x) = \{i \in D_1 : B(H(X_i)) = B(H(x))\}$. Since binning procedures are in-sample auto-calibrated, the predictive system Π^{BIN} contains an *auto-calibrated* predictive CDF for Y_{n+1} (Allen et al., 2025). As a consequence, it also implies probabilistic/threshold/quantile calibration (Gneiting and Resin, 2023).

Isotonic calibration: CIDR. We apply conformal isotonic distributional regression (CIDR) (Allen et al., 2025) using a covariate summary $H(x)$ together with an order $\preceq$ on $H(\mathcal{X})$, yielding a procedure G^{IDR} that is isotonically calibrated in-sample. Since IDR is most naturally formulated in terms of survival functions (so the order aligns with the required monotonicity and admits a universal estimator for infinite thresholds), we work with the survival representation throughout. Concretely, when the base forecaster provides values on fixed thresholds $t_1, \dots, t_m$, we take $H(x) = (\hat{S}(t_1 \mid x), \dots, \hat{S}(t_m \mid x)) \in [0,1]^m$ with $\hat{S}(t \mid x) = 1 - \hat{F}(t \mid x)$, and order this vector together with the ones in the calibration set coordinate-wise: $u \preceq v$ iff $u_k \le v_k$ for all k. To mitigate the prevalence of incomparable pairs in high dimension, we optionally replace the full vector by a lower-dimensional summary, e.g. $H(x) = (\hat{S}(t_{j_5} \mid x), \hat{S}(t_{j_{10}} \mid x), \hat{S}(t_{j_{15}} \mid x))$ for a few thresholds; this is the index-model route advocated for conformal IDR (Allen et al., 2025). Therefore, the predictive system Π^{IDR} contains an isotonically calibrated predictive survival distribution for Y_{n+1} (Allen et al., 2025).

Grid calibration: CVAPS. We propose conformal Venn-Abers predictive systems (CVAPS) for grid calibration on a fixed threshold set $T = \{t_1 < \cdots < t_m\}$. A predictive CDF $F(\cdot \mid X)$ is *grid-calibrated on* T if for every $k \in \{1, \dots, m\}$,

$$\mathbb{E}[\mathbb{1}\{Y \le t_k\} \mid F(t_k \mid X)] = F(t_k \mid X) \quad \text{a.s.}$$

Equivalently, each binary event $Z_k = \mathbb{1}\{Y \le t_k\}$ is calibrated with respect to the predicted probability $F(t_k \mid X)$ (perfect reliability at each grid threshold). This notion is weaker than threshold, isotonic, and auto-calibration: each implies grid calibration on any T, but not conversely.

CVAPS calibrates each threshold event separately using split Venn-Abers calibration (Vovk and Petej, 2014) applied to the base score $s_k(x) := \hat{F}(t_k \mid x)$ with labels $z_{ik} = \mathbb{1}\{Y_i \le t_k\}$, yielding an interval-valued prediction $[\underline{p}_k(x), \overline{p}_k(x)]$ for the calibrated non-exceedance probability at t_k. Stacking these intervals across k produces pointwise lower/upper bands on T, but treating thresholds independently can violate CDF monotonicity and yield decreasing bounds (Nouretdinov et al., 2018). To enforce monotonicity *without losing the per-threshold Venn-Abers containment*, we take the monotone hull that *expands* (never shrinks) the original intervals:

$$\Pi_u(t_k \mid x) := \max_{j \le k} \overline{p}_j(x), \quad \Pi_\ell(t_k \mid x) := \min_{j \ge k} \underline{p}_j(x).$$

This yields nondecreasing bounds and preserves per-threshold containment since $\Pi_\ell(t_k \mid x) \le \underline{p}_k(x) \le$

Table 1: Performance on the bimodal extreme synthetic experiment (mean $[q_{25}, q_{75}]$). We group metrics by (i) full-distribution quality (CRPS, dispersion), (ii) 95% prediction-interval (PI) quality (coverage, efficiency), and (iii) threshold calibration (ECE at selected quantiles).

	Full distribution		**95% PI**		**Threshold calibration**
Method	CRPS	Dispersion	Coverage	Efficiency	ECE
Uncalibrated	0.092 [0.087, 0.096]	0.085 [0.078, 0.091]	0.942 [0.920, 0.970]	0.523 [0.496, 0.551]	0.044 [0.035, 0.051]
CMCPS	0.091 [0.086, 0.096]	0.081 [0.075, 0.086]	0.953 [0.940, 0.970]	0.538 [0.487, 0.583]	0.040 [0.032, 0.046]
CBIN	0.092 [0.086, 0.098]	0.072 [0.067, 0.076]	0.994 [0.990, 1.000]	0.934 [0.898, 0.979]	0.067 [0.058, 0.075]
CIDR	0.223 [0.228, 0.247]	0.016 [0.005, 0.021]	0.992 [0.990, 1.000]	0.909 [0.902, 0.977]	0.048 [0.030, 0.061]
CVAPS	0.090 [0.084, 0.095]	0.068 [0.063, 0.072]	0.968 [0.950, 0.980]	0.549 [0.520, 0.581]	0.041 [0.034, 0.047]

$\overline{p}_k(x) \leq \Pi_u(t_k \mid x)$. Finally, note that Venn-Abers calibration can be viewed as a special case of CIDR with a binary label (Allen et al., 2025).

4 Synthetic Experiment

We evaluate whether the R2C discrete CDF forecaster (Section 2) can represent a challenging conditional law (bimodality + heteroskedasticity) and how the calibrators from Section 3 trade off *finite-sample reliability* and *sharpness*.

We propose a slightly adjusted experiment of Lei and Wasserman (2014). We sample $X \sim \mathrm{Unif}([-1.5, 1.5])$ and generate

$$Y \mid X = x \sim \tfrac{1}{2}\mathcal{N}\big(f(x) - g(x),\, \sigma(x)^2\big) + \tfrac{1}{2}\mathcal{N}\big(f(x) + g(x),\, \sigma(x)^2\big),$$

where $f(x) = (x-1)^2(x+1)$, $g(x) = 4(x+0.5)\mathbb{1}\{x \geq -0.5\}$, and $\sigma(x) = \sqrt{\frac{1}{4} + |x|}$.

We compare the uncalibrated R2C baseline (PMF head $\to$ monotone grid-CDF, trained with JBCE) against CM-CPS, CBIN, CIDR, and CVAPS (Section 3) using a split setup (train base model; fit calibration; evaluate on an independent test set). We repeat the experiment 100 times with $n_{\text{train}} = 100$, $n_{\text{cal}} = 300$, and $n_{\text{test}} = 100$ (using a small training set to avoid near-perfect generalization on this simple benchmark). We use $m = 50$ equal-width bins (threshold grid T). The baseline outputs a single CDF $\hat{F}(\cdot \mid x)$, while CPS methods output a predictive system band $\Pi(\cdot \mid x) = [\Pi_\ell(\cdot \mid x), \Pi_u(\cdot \mid x)]$ (Allen et al., 2025). For single-CDF metrics (continuous ranked probability score (CRPS), dispersion of PIT), we use the worst-case-CRPS minimizer over the band (Allen et al., 2025):

$$\widehat{F}(y \mid x) = \Pi_u(y \mid x) - \tfrac{1}{2}\Pi_u(y \mid x)^2 + \tfrac{1}{2}\Pi_\ell(y \mid x)^2.$$

We report (i) CRPS and dispersion, (ii) 95% PI coverage and efficiency (mean width), and (iii) threshold calibration via an expected calibration error (ECE) and CORP thresholded reliability diagrams (Dimitriadis et al., 2021). For the baseline, the 95% prediction interval (PI) is obtained by inverting $\hat{F}$ on the grid; for CPS methods we use the conservative band inversion (valid for any CDF consistent with Π): for $\alpha = 0.05$, $I_\alpha(x) = \big[\inf\{y : \Pi_u(y \mid x) \geq \alpha/2\}, \inf\{y : \Pi_\ell(y \mid x) \geq 1 - \alpha/2\}\big]$. For CIDR, we use the index summary $H(x) = (\hat{S}(t_{12} \mid x), \hat{S}(t_{25} \mid x), \hat{S}(t_{38} \mid x))$ (survival values at three grid thresholds), and for CBIN we apply k-means to $H(X)$ with 20 clusters. See Table 1 and Figs. 1-2.

Results. The uncalibrated R2C model already captures the multi-modal and heteroskedastic structure of the conditional law (Fig. 1), but it mildly undercovers at the 95% level (Table 1). CM-CPS moves coverage toward nominal and improves PIT behavior with only small changes in CRPS/dispersion. Across methods, CVAPS provides the most favorable overall trade-off on this benchmark: it yields strong threshold calibration on the grid (low ECE), competitive (often improved) CRPS/dispersion, and near-nominal (slightly conservative) PI coverage (a natural consequence of constructing the bands considering the upper and lower bounds), while producing tight CDF envelopes that track the ground-truth well (Fig. 1). In contrast, CBIN and CIDR become conservative here: both achieve very high coverage (close to 1) but at the cost of much wider intervals (low efficiency) and, for CIDR, substantially worse CRPS, consistent with the difficulty of stable multivariate isotonic constraints when CDF features cross under heteroskedasticity/multimodality.

5 Discussion

Our results illustrate a trade-off between the strength of the calibration notion and the informativeness of the resulting predictive distributions. Methods targeting stronger conditional notions (e.g., auto-calibration via binning, isotonic constraints in CIDR) can become conservative when the conditional law varies rapidly

with x or when high-dimensional CDF features are frequently incomparable due to heteroskedasticity and multimodality. In contrast, grid calibration (CVAPS) leverages the R2C grid directly and appears to deliver a favorable balance of reliability and sharpness in our experiments. A limitation of our current CIDR implementation is the sensitivity of multivariate isotonic regression to the chosen covariate representation and partial order.

Acknowledgements

Jef Jonkers is funded by the Research Foundation Flanders (FWO, Ref. 1S11525N). Part of the research was funded through the Research Foundation Flanders senior research project on Trustworthy Time-to-Event Predictions (FWO, Ref. G0AH525N). Part of this research was supported through the Flemish Government (AI Research Program).

References

Allen, S., Gavrilopoulos, G., Henzi, A., Kleger, G.-R., and Ziegel, J. (2025). In-sample calibration yields conformal calibration guarantees. arXiv:2503.03841 [stat].

Berg, A., Oskarsson, M., and O'Connor, M. (2021). Deep Ordinal Regression with Label Diversity. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2740–2747.

Chen, B., Islam, M., Gao, J., and Wang, L. (2022). Deconvolutional Density Network: Modeling Free-Form Conditional Distributions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6183–6192. Number: 6.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8):e2016191118.

Foresi, S., , and Peracchi, F. (1995). The Conditional Distribution of Excess Returns: An Empirical Analysis. *Journal of the American Statistical Association*, 90(430):451–466.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). Deep Ordinal Regression Network for Monocular Depth Estimation. pages 2002–2011.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T. and Resin, J. (2023). Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2):3226–3286.

Guha, E. K., Natarajan, S., Möllenhoff, T., Khan, M. E., and Ndiaye, E. (2023). Conformal Prediction via Regression-as-Classification.

Jonkers, J., Coopman, F., Duchateau, L., Van Wallendael, G., and Van Hoecke, S. (2026). Reliable uncertainty quantification for 2D/3D anatomical landmark localization using multi-output conformal prediction. *Medical Image Analysis*, 110:103953.

Jonkers, J., Verhaeghe, J., Wallendael, G. V., Duchateau, L., and Hoecke, S. V. (2025). Conformal Convolution and Monte Carlo Meta-learners for Predictive Inference of Individual Treatment Effects. arXiv:2402.04906 [cs].

Lei, J. and Wasserman, L. (2014). Distribution-free Prediction Bands for Non-parametric Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96.

Li, R., Reich, B. J., and Bondell, H. D. (2021). Deep distribution regression. *Computational Statistics & Data Analysis*, 159:107203.

Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S. (2014). Heteroscedastic Extended Logistic Regression for Postprocessing of Ensemble Guidance.

Nouretdinov, I., Volkhonskiy, D., Lim, P., Toccaceli, P., and Gammerman, A. (2018). Inductive Venn-Abers predictive distribution. In *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, pages 15–36. PMLR.

Pintea, S. L., Lin, Y., Dijkstra, J., and Van Gemert, J. C. (2023). A step towards understanding why classification helps regression. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19915–19924, Paris, France. IEEE.

Stewart, L., Bach, F., Berthet, Q., and Vert, J.-P. (2023). Regression as Classification: Influence of Task Formulation on Neural Network Features. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 11563–11582. PMLR.

Tansey, W., Pichotta, K., and Scott, J. G. (2016). Better Conditional Density Estimation for Neural Networks. arXiv:1606.02321 [stat].

van den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel Recurrent Neural Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1747–1756. PMLR.

Vovk, V. and Petej, I. (2014). Venn-Abers predictors. arXiv:1211.0025 [cs, stat].

Vovk, V., Petej, I., Toccaceli, P., Gammerman, A., Ahlberg, E., and Carlsson, L. (2020). Conformal calibrators. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, pages 84–99. PMLR.

Vovk, V., Shen, J., Manokhin, V., and Xie, M.-G. (2019). Nonparametric predictive distributions based on conformal prediction. *Machine Language*, 108(3):445–474.

Weigend, A. S. and Srivastava, A. N. (1995). Predicting conditional probability distributions: a connectionist approach. *International Journal of Neural Systems*, 06(02):109–118.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16(3):361–368.

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful Image Colorization. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham. Springer International Publishing.

A Additional figures & tables

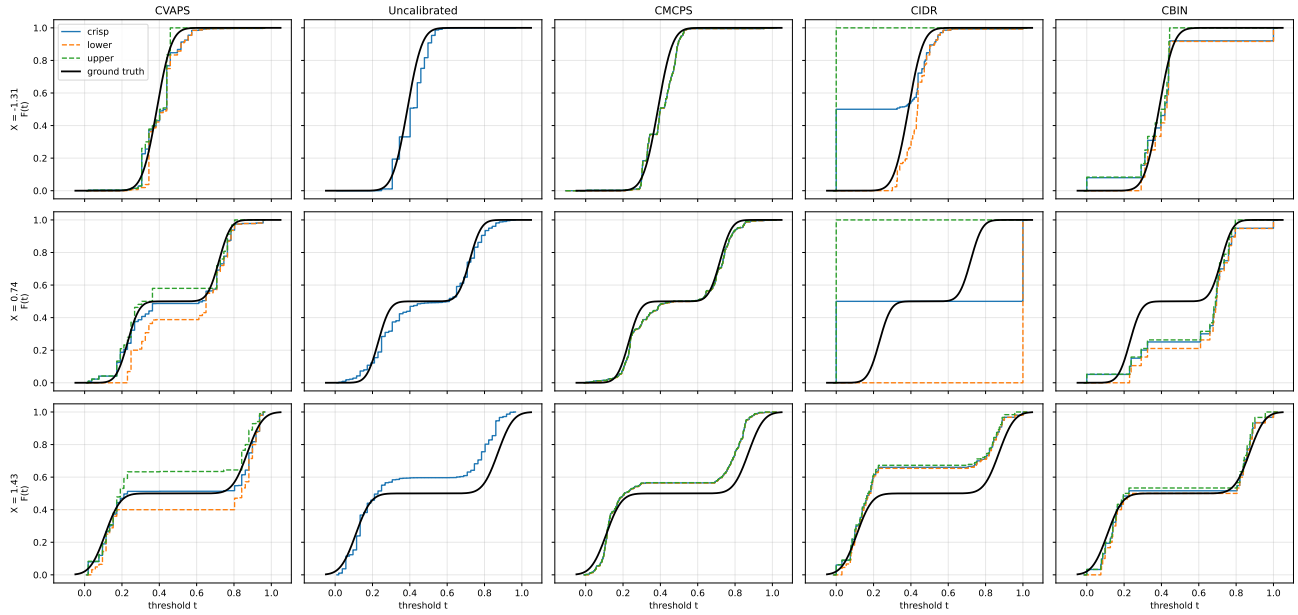

Figure 1: Conditional CDF estimates on the bimodal extreme synthetic DGP. For three test inputs x (rows), we plot the predicted empirical CDF $\hat{F}_{Y|X=x}(t)$ for five methods (columns: CVAPS, uncalibrated, CMCPS, CIDR, CBIN). Dashed curves denote lower/upper CDF bounds when available, and the black curve is the ground-truth conditional CDF computed from the known mixture model.

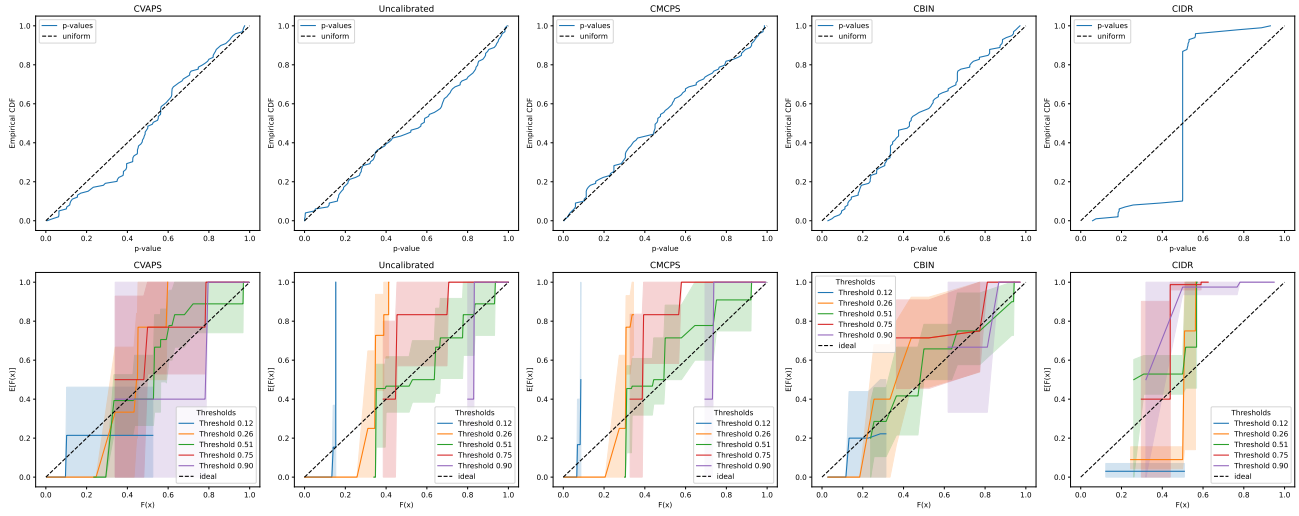

Figure 2: Calibration diagnostics on the bimodal extreme synthetic experiment. Top row: probability integral transform (PIT) plots, showing the empirical CDF of randomized PIT values versus the uniform reference (dashed). Bottom row: CORP thresholded reliability diagrams (Dimitriadis et al., 2021) at five response quantiles (10%, 25%, 50%, 75%, 90%), plotting the empirical event frequency $\mathbb{E}[\mathbb{1}\{Y \leq t\}]$ against the predicted probability $\hat{F}_{Y|X}(t)$; shaded bands denote 95% confidence intervals and the dashed line indicates perfect calibration